

MULTIMEDIA



UNIVERSITY

STUDENT ID NO

--	--	--	--	--	--	--	--	--	--

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 1, 2018/2019

TMM 3341– DATA MINING AND MACHINE LEARNING SYSTEMS

(All sections / Groups)

22 OCTOBER 2018

2.30pm – 4.30pm

(2 Hours)

INSTRUCTIONS TO STUDENT

1. This Question paper consists of 4 pages with 5 Questions only.
2. Answer all **FIVE** questions. All questions carry equal marks and the distribution of the marks for each question is given.
3. Please write all your answers in the Answer Booklet provided.

QUESTION 1

- a. Explain briefly the **TWO (2)** main tasks of Data Mining. (4 marks)
- b. Explain briefly why data cleaning is a necessary task for data mining. (3 marks)
- c. An internet provider company realized that in recent years it has been losing several of its long-standing customers (more than 3 years) due to competitors. In order to fight this phenomenon, the company needs to identify potential churners and also understand what makes (or will make) them leave. Briefly describe a project plan to help the company, using data mining methodology. Clearly remark the choices and assumptions made in the process.

The company has the following information:

- All demographic information on its customers.
- The details of all their transactions (calls, data traffic usage, etc.) and contracts (including type of contract, start date and termination).

(3 marks)

QUESTION 2

- a. Ali, owner of Fast Food is interested to find out the correlation between his two most sold products namely FishChips and Satay. Ali analyzes his data base and finds the following statistics about the two items.

Table 1: Contingency Table

	FishChips	$\overline{\text{FishChips}}$	Σ_{row}
Satay	6000	1500	7500
$\overline{\text{Satay}}$	3000	4500	7500
Σ_{col}	9000	6000	15000

- i. Suppose that the association rule buys (X, "FishChips") $\Rightarrow$ buys(X, "Satay") is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, determine whether this rule has strong association? (3 marks)
- ii. Based on the given data, evaluate whether the purchase of FishChips is independent of the purchase of satay? If not, what kind of correlation relationship exists between the two? (3 marks)
- b. Define *market basket analysis*. Describe **TWO (2)** methods on how this analysis can help retailers. (4 marks)

Continued

QUESTION 3

- a. Given a group of data: 800, 2000, 400, 2200, 1000, 200
- Transform the group of data onto the range [0, 100] using min-max normalization. (3 marks)
 - For the above group of data, partition them into two bins using equal-width partitioning method and equal-frequency partitioning. Show your steps. (3 marks)
- b. Explain the difference between *machine learning* and “*normal programming*”? (4 marks)

QUESTION 4

- a. Using the dataset shown in Table 2, classify data point $x=5.0$ according to its 1, 3, 5, and 9-nearest neighbors (using majority vote). (3 marks)

Table 2: Data Points

x	0.5	3.0	4.4	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	Bad	Bad	Good	Good	Good	Bad	Bad	Good	Bad	Bad

- b. Given the training set (Table 3a), build a Naive Bayes classification model and apply it to the test set (Table 3b). (5 marks)

Table 3a: Training Data

Outlook	Temp	Windy	class
high	no	false	Y
medium	no	true	Y
low	yes	false	N
high	no	true	N
low	yes	true	Y
high	no	false	Y
medium	yes	false	N

Table 3b: Test Data

Outlook	Temp	Windy	class
low	no	false	
high	yes	true	
medium	yes	true	

- c. Discuss **ONE** (1) advantage of *agglomerative* hierarchical clustering algorithm. (2 marks)

Continued

QUESTION 5

- a. Are the two clusters shown in Figure 1 well separated? Justify your answer. (2 marks)

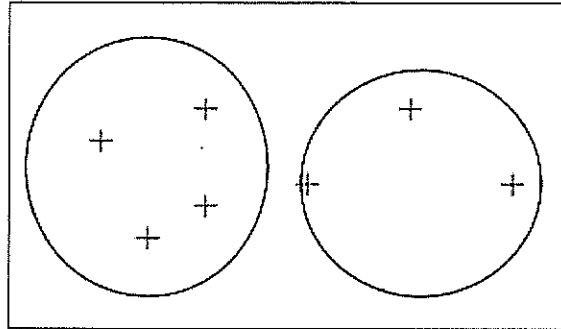


Figure 1: Data Points

- b. Explain the difference between overfitting and under-fitting. Why overfitting or under-fitting is “undesirable” in machine learning? (3 marks)
- c. Present an example where data mining is crucial to the success of a business. What data mining function does this business need? (5 marks)

End of Page